Pattern Analysis in Smart City Mobility - An Application in Singapore

CIKM AnalytiCup in DataSpark Mobility - Finalist Report

Camelia Elena Ciolac
ciolac_c@yahoo.co.uk

ABSTRACT

This paper presents the implemented solution for the CIKM AnalytiCup in DataSpark Mobility Open-Task Challenge. Using publicly available APIs we have built a system able to extract both recurrent and unsystematic patterns of mobility from data registered in Singapore in terms of human mobility. The implemented data architecture combines Big data and data science technologies with the aim to discover the rhythm of the city.

KEYWORDS

mobility, smart city, pattern analysis

ACM Reference format:
Camelia Elena Ciolac. 2017. Pattern Analysis in Smart City Mobility - An Application in Singapore. In *Proceedings of CIKM AnalytiCup 2017-DataSpark Mobility, 2017, Singapore,* 4 pages.
DOI: 10.475/123_4

1 PROBLEM STATEMENT

Nowadays, more and more cities roll out testbeds for the Internet of Things across districts (e.g. ”iCITY” in the Olympic Park in London) or even the entire city (e.g. ”City of Things” in Antwerp - Belgium, ”SmartSantander” in Santander - Spain). While useful for monitoring environmental parameters (e.g. temperature, humidity, noise, air quality, light), road traffic congestion and utilities supply (e.g. water, gas, electricity), the data collected from these sensors has limited ability to tell the story of the city’s daily life.

The task that we proposed to address is the identification of patterns in human mobility within a large and densely populated city such as Singapore. We aim to understand the city’s rhythm.

From the current state of the art in the literature on this topic, three main practical approaches are spread the most. The first is based on tracking volunteered individuals (through crowd-sourcing mobile apps or through on-board GPS loggers in private cars/taxis), thus operating on collections of individual GPS traces. In the second approach, call detail records (CDR) from telecom providers are used. However, rarely is the case that the cellular network cells follow the shapes of administrative subzones and thus frequently this relation is often overlooked. The third approach is to use geo-tagged social media (e.g. Tweets, Instagram photos, FourSquare check-ins and others) to infer the travel patterns and trajectories in town.

2 OVERALL ARCHITECTURE

As described in the next sections, the application combines multiple technologies for ingesting, managing and analyzing the data. An overview of the architecture is depicted in Figure 1.

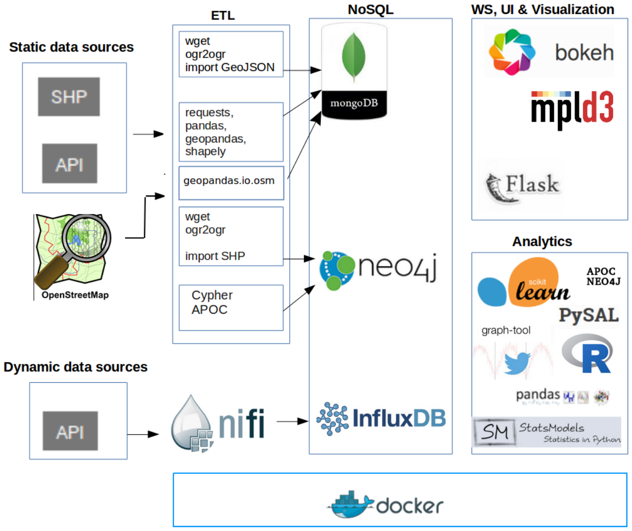


Figure 1: Overall solution architecture

3 DATA SOURCES AND DATA COLLECTION

The data sources used in this application comprised:

- DataSpark APIs [1]: footfall and origin-destination matrix;
- Environmental data [2] : air temperature, rainfall and wind speed;
- e-Government open data [3]: residents by subzone age group and gender (june 2016), residents by subzone and type of dwelling (june 2016), historic sites, museums, monuments, park facilities, tourist attractions;
- OpenStreetMap data [4]: buildings, offices, amenities, shops, leisure, cultural and tourism places.
- Public bus transport network (LTA DataMall) [5]: bus stops, bus routes.

Historic data was collected and preprocessed in batch, however for the live application, data is ingested using dataflows developed

CIKM AnalytiCup 2017- DataSpark Mobility, 2017

DOI: 10.475/123_4

[1] https://apistore.datasparkanalytics.com:8243/
[2] https://api.data.gov.sg/v1/environment/
[3] https://data.gov.sg/dataset/
[4] https://www.openstreetmap.org
[5] http://datamall2.mytransport.sg/ltaodataservice

in Apache NiFi. This choice of technology was based on specific benefits of Apache NiFi in terms of development simplicity (being configuration-based), queues on all inter-processor communications to handle back-pressure, data provenance tracking as objects flow through the system, timer-based triggering of components' execution, JOLT for JSON manipulation, connectors to NoSQL.

4 DATA MODELLING AND MANAGEMENT

Dealing with this variety of data formats and models lead us to opt for a polyglot data store. It consists of:

- a document-oriented database in MongoDB which stores spatial and structured data about Singapore subzones;
- a graph database in Neo4j which stores spatial data about Singapore subzones and bus transportation network - in the form of vertices and edges;
- a timeseries database in InfluxDB which stores the timeseries corresponding to footfall and OD matrix (from DataSpark APIs) as well as the environment recordings from the weather APIs.

For the various functionalities described in the following sections, data from one or more of these databases were integrated. For example, when computing complex networks metrics using the recorded OD matrix, we took advantage of Neo4j's APOC to select data from InfluxDB through its REST endpoint and then create temporary "od" edges between the subzone nodes.

Neo4j spatial plugin was used for many spatial integrations of the datasets, an example being depicted in Figure 2, where the bus stop geo-points were intersected with the subzones polygons to create "inside" relationships. Multiple edges connect pairs of BUSSTOP vertices, one for each bus route. The same spatial plugin was used for creating relationships between subzones nodes which were geographically adjacent.

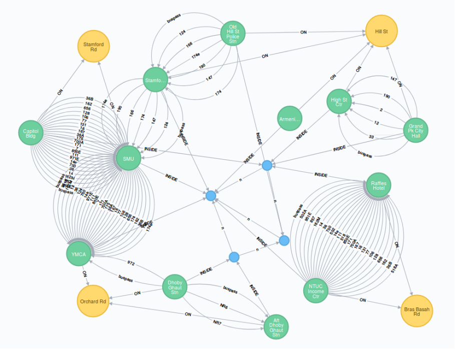


Figure 2: Subgraph from Neo4j, showing subzones(blue), busstops(green), roads(yellow) and their relationships

As for data modeling in InfluxDB, data from each DataSpark API is stored in a distinct measurement. For discrete visits, data for each subzone is stored in a separate field and the tags hold the API request parameters (e.g. time granularity, filter dimensions) and metric (e.g. footfall, total_visits). Instead, for OD matrix data, we modelled to have a single field named "value" and to keep in tags the origin_subzone and destination_subzone - which gives more flexibility in filtering. Moreover, InfluxDB demonstrated many benefits in window aggregations, at different time granularities, and continuous queries.

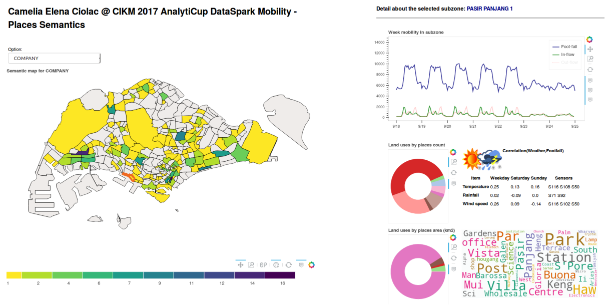
Derived data, resulted after specific processing on raw mobility data, are stored in MongoDB (e.g. correlation weather-footfall, nearest weather sensors per subzone,cluster id from K-means clustering) as well as in InfluxDB measurements (e.g. the anomalies and the structural changes identified in the footfall timeseries of each subzone).

5 DATA ANALYTICS

In this section we walk through the functionality implemented.

5.1 Places Semantics

To begin with, places semantics' were analyzed from multiple data sources. On one hand, demographics and dwelling were extracted from official sources. They were complemented, for the purpose of this study, with data extracted from OpenStreetMap. The latter served as a source to access building-level information and thus allowed computing land use on specific topics of interest (e.g. universities, healthcare, industrial areas, company offices, shopping malls). A wordcloud with tags extracted from all bus stop descriptions, per subzone, was computed and added in order to support the understanding of places' semantics.



Camelia Elena Ciolac @ CIKM 2017 AnalytiCup DataSpark Mobility - Places Semantics

Option: COMPANY

Semantic map for COMPANY

Detail about the selected subzone: PASIR PANJANG 1

Correlation(Weather,Footfall)

Item	Weekday	Saturday	Sunday	Sensors
Temperature	0.25	0.13	0.16	S116 S108 S50
Rainfall	0.02	-0.09	0.0	S71 S92
Wind speed	0.26	0.09	-0.14	S116 S102 S50

Figure 3: Places semantics

Further on, weather data was collected and analyzed to identify, per subzone, the correlation between temperature, rainfall and wind speed - on one hand- and footfall. The analysis was performed separately for workdays, for Saturdays and for Sundays.

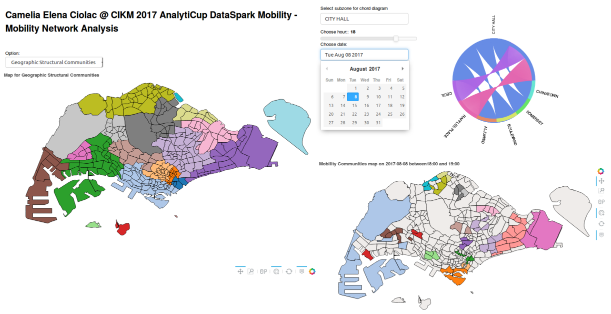
The data visualization is organized in this screen based on a master-detail policy. The choropleth map on the left side of the web page displays the spatial distribution of topics of interest. Once a subzone becomes interesting for its contents (e.g. leads the top in the number of companies, such as Raffles Place or Kian Teck), the user can select it on the map and obtain detailed information about it on the right side of the webpage. This information covers: a sample week of hourly mobility data (footfall, in-flows and outflows), the composition of land use (in counts and in surface), the wordcloud and the weather - mobility relationship for that subzone.

In this manner, the user observes the difference in the weekly mobility pattern (between the two aforementioned subzones) and can judge it aided by the difference in land use (e.g. Kian Teck has a significant residential and industrial scope, whereas Raffles Place has more entertainment spaces).

5.2 Complex Networks

Multiple metrics exist in Complex Networks theory to characterize the structure of a graph. We selected some of them and computed them for static data and for dynamic data. The 8 analysis are:

- Geography-based Betweenness centrality
- Geography-based Closeness centrality
- Geography-based PageRank importance
- Geography-based Community detection
- Bus network Degree centrality
- Bus network Closeness centrality
- Bus network PageRank importance
- OD matrix Community detection (hourly)


Camelia Elena Ciolac @ CIKM 2017 AnalytiCup DataSpark Mobility - Mobility Network Analysis

Option: Geographic Structural Communities

Map for Geographic Structural Communities

Select subzone for chord diagram: CITY HALL

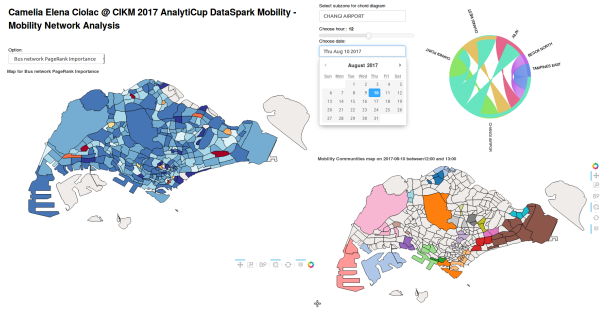
Choose hour:: 18

Choose date: Tue Aug 08 2017

Mobility Communities map on 2017-08-08 between18:00 and 19:00

Figure 4: Complex networks metrics - geography-based and mobility-based communities in the subzones graph

The geography-based metrics calculation takes into account the graph resulted from modeling subzones' spatial adjacency relationships. The bus network based metrics calculation takes into account the graph of busstops linked by bus routes. The OD matrix based metrics calculation uses the graph induced by the mobility between subzones at different times in different days of the calendar year.


Camelia Elena Ciolac @ CIKM 2017 AnalytiCup DataSpark Mobility - Mobility Network Analysis

Option: Bus network PageRank Importance

Map for Bus network PageRank Importance

Select subzone for chord diagram: CHANGI AIRPORT

Choose hour:: 12

Choose date: Thu Aug 10 2017

Mobility Communities map on 2017-08-10 between12:00 and 13:00

Figure 5: Complex networks metrics - bus stops Page Rank importance and dynamic communities

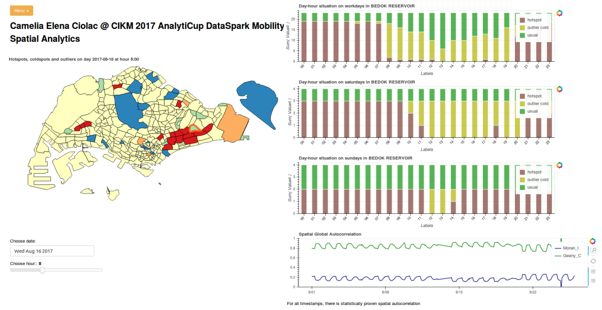
Besides comparing dynamic and static results of the subzones graph structure, the analysis can also be carried out along the temporal axis to observe changes in subzones' communities structure in terms of mobility between time intervals of a day or between different days.

5.3 Spatial Analytics

Given the hourly footfall at subzone level during the day, we tested if there exists spatial autocorrelation. Spatial autocorrelation means non-random pattern of attribute values over the set of spatial units, the subzones. It resulted that the null hypothesis of a random process operating in space can be rejected, as both Moran's I and Geany's C indicators indicated positive spatial autocorrelation. This means that similar values of traffic tend to be close in space. The choropleth maps support visually the findings.

Moreover, we computed for each hour of the day the hotspots (high footfall subzones surrounded by high footfall neighbor subzones) and coldspots (low values of footfall surrounded by low values of footfall).

Spatial outliers were identified based on this criterion to be subzones with footfall numbers completely opposite from neighbors (e.g. low-high or high-low). An example of the high-low outlier is Changi Airport subzone.


Camelia Elena Ciolac @ CIKM 2017 AnalytiCup DataSpark Mobility Spatial Analytics

Hotspots, coldspots and outliers on day 2017-08-16 at hour 8:00

Choose date: Wed Aug 16 2017

Choose hour:: 8

For all timestamps, there is statistically proven spatial autocorrelation

Figure 6: Spatial analytics - spatial autocorrelation

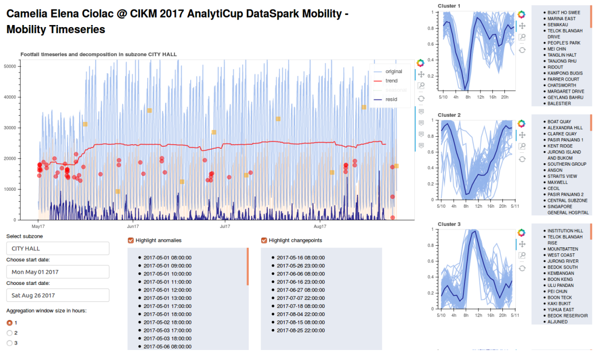
Secondly, we applied various thresholds and used Join counts to verify if the spatial autocorrelation still holds on this binary data. Depending on the date-time and on the chosen threshold, we observed both scenarios - dispersion and random pattern.

5.4 Temporal Analytics

In the time domain, it is of interest to analyze subzones individually and in their entire set.

Considering all subzones for a chosen period of time and window size, we first aggregated the footfall. A window size of 1 hour left the series in their raw form. We then performed timeseries decomposition and did K-means clustering on a cycle of the seasonal component scaled to [0,1]. In this manner we accomplished the objective of finding structural groups in the repetitive component of the timeseries, not influenced by differences in values magnitude. The focus was strictly on the traffic profile.

As somehow expected, the highly populated residential areas (Bedok North/South, Tampines East/West and others) were clustered together, whereas the central area (City Hall, Boat Quay, Chinatown, Raffles Place and others) were clustered together too.



Camelia Elena Ciolac @ CIKM 2017 AnalytiCup DataSpark Mobility - Mobility Timeseries

Footfall timeseries and decomposition in subzone CITY HALL

Select subzone

CITY HALL

Choose start date:

Mon May 01 2017

Choose start date:

Sat Aug 26 2017

Aggregation window size in hours:

● 1
○ 2
○ 3

☑ Highlight anomalies

- 2017-05-01 08:00:00
- 2017-05-01 09:00:00
- 2017-05-01 10:00:00
- 2017-05-01 11:00:00
- 2017-05-01 12:00:00
- 2017-05-01 13:00:00
- 2017-05-01 17:00:00
- 2017-05-01 18:00:00
- 2017-05-02 18:00:00
- 2017-05-03 17:00:00
- 2017-05-03 18:00:00
- 2017-05-06 08:00:00

☑ Highlight changepoints

- 2017-05-16 08:00:00
- 2017-05-26 23:00:00
- 2017-06-06 08:00:00
- 2017-06-16 23:00:00
- 2017-06-27 08:00:00
- 2017-07-07 22:00:00
- 2017-07-18 08:00:00
- 2017-08-04 22:00:00
- 2017-08-15 08:00:00
- 2017-08-25 22:00:00

Cluster 1

- BUKIT HO SWEE
- MARINA EAST
- SEMAKAU
- TELOK BLANGAH DRIVE
- PEOPLE'S PARK
- MEI CHIN
- TANGLIN HALT
- TANJONG RHU
- RIDOUT
- KAMPONG BUGIS
- FARRER COURT
- CHATSWORTH
- MARGARET DRIVE
- GEYLANG BAHRU
- BALESTIER

Cluster 2

- BOAT QUAY
- ALEXANDRA HILL
- CLARKE QUAY
- PASIR PANJANG 1
- KENT RIDGE
- JURONG ISLAND AND BUKOM
- SOUTHERN GROUP
- ANSON
- STRAITS VIEW
- MAXWELL
- CECIL
- PASIR PANJANG 2
- CENTRAL SUBZONE
- SINGAPORE GENERAL HOSPITAL

Cluster 3

- INSTITUTION HILL
- TELOK BLANGAH RISE
- MOUNTBATTEN
- WEST COAST
- JURONG RIVER
- BEDOK SOUTH
- KEMBANGAN
- BOON KENG
- ULU PANDAN
- PEI CHUN
- BOON TECK
- KAKI BUKIT
- YUHUA EAST
- BEDOK RESERVOIR
- ALJUNIED

Figure 7: Temporal analytics

At the individual level of a subzone, the interest is on finding anomalies in the timeseries and change points in its trend. For this we used R libraries recently open sourced by Twitter for detecting anomalies and breakouts. The anomaly results in footfall are highlighted in the timeseries plot (red and orange markers) as shown in Figure 7.

6 POTENTIAL IMPACT OF SOLUTION

The target is to facilitate a better understanding of the mobility phenomena for future urbanistic development.

The analysis so far brought many interesting insights into the daily life of the city, its rhythm on workdays/week-ends and holidays, the similarity and dissimilarity between neighboring subzones in space with respect to traffic, the shifts in temporal trend and anomalies in mobility timeseries.

Following the methodology that we've employed, this solution can be generalized and applied in other major cities, too.

7 REFERENCES

[1] Richard Becker, Ramon Caceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. 2013. **Human mobility characterization from cellular network data.** Commun. ACM 56, 1 (January 2013), 74-82.
DOI: https://doi.org/10.1145/2398356.2398375

[2] Dan Tasse, Jason I Hong. 2014. **Using Social Media to Understand Cities**. http://repository.cmu.edu/cgi/viewcontent.cgi?article=1271&context=hcii

[3] Fabio Miranda, Harish Doraiswamy Member, Marcos Lage, Kai Zhao, Bruno Goncalves, Luc Wilson, Mondrian Hsieh, and Claudio T. Silva. **Urban Pulse: Capturing the Rhythm of Cities**. http://www.bgoncalves.com/download/finish/4/300.html

[4] R. A. Becker et al.2011. **A Tale of One City: Using Cellular Network Data for Urban Planning**, in IEEE Pervasive Computing, vol. 10, no. 4, pp. 18-26, April 2011.doi: 10.1109/MPRV.2011.44

[5] Paola Pucci, Fabio Manfredini, Paolo Tagliolato. 2015. **Mapping Urban Practices Through Mobile Phone Data**. Springer, 2015

[6] Francesco Calabrese, Laura Ferrari, Vincent Blondel. 2014. **Urban Sensing Using Mobile Phone Network Data: A Survey of Research**. http://researcher.watson.ibm.com/researcher/files/ie-FCALABRE/survey.pdf

[7] Sebastian Grauwin, Stanislav Sobolevsky, Simon Moritz, Istvan Godor, Carlo Ratti.2014. **Towards a comparative science of cities: using mobile traffic records in New York, London and Hong Kong**. https://arxiv.org/pdf/1406.4400.pdf

[8] Andras Garzo, Andras A. Benczur, Csaba Istvan Sidlo, Daniel Tahara, Erik Francis Wyatt. **Real-time streaming mobility analytics**. http://eprints.sztaki.hu/7657/1/Garzo_697_2476911_ny.pdf

[9] Anastasios Noulas, Cecilia Mascolo, Enrique Frias-Martinez.2013. **Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments.** https://www.cl.cam.ac.uk/~cm542/papers/mdm2013.pdf

[10] Chen Zhong, Stefan Muller Arisona, Xianfeng Huang, Michael Batty, Gerhard Schmitt. 2014. **Detecting the dynamics of urban structure through spatial network analysis**,International Journal of Geographical Information Science, DOI: 10.1080/13658816.2014.914521